Preprint 2025a CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale.
Zhun Wang*, Tianneng Shi*, Jingxuan He, Matthew Cai, Jialin Zhang, Dawn Song.
Adopted in Anthropic's Claude Sonnet 4.5 System Card.
Top 0.6% of submissions at ICLR 2026 (under review).

ICML 2025a BaxBench: Can LLMs Generate Secure and Correct Backends?
Mark Vero, Niels Mündler, Victor Chibotaru, Veselin Raychev, Maximilian Baader, Nikola Jovanović, Jingxuan He, Martin Vechev.
Spotlight Paper.

CCS 2023 Large Language Models for Code: Security Hardening and Adversarial Testing.
Jingxuan He, Martin Vechev.
Distinguished Paper Award.

PLDI 2025 Type-Constrained Code Generation with Language Models.
Niels Mündler*, Jingxuan He*, Hao Wang, Koushik Sen, Dawn Song, Martin Vechev.
Featured as #1 on Hacker News.